Structured Adversarial Training for Unsupervised Monocular Depth Estimation

Ishit Mehta Parikshit Sakurikar P J Narayanan
Center for Visual Information Technology, Kohli Center on Intelligent Systems
International Institute of Information Technology, Hyderabad, India
{ishit.mehta, parikshit.sakurikar}@research.iiit.ac.in, pjn@iiit.ac.in

Abstract

The problem of estimating scene-depth from a single image has seen great progress lately. Recent unsupervised methods are based on view-synthesis and learn depth by minimizing photometric reconstruction error. In this paper, we introduce Structured Adversarial Training (StrAT) to this problem. We generate multiple novel views using depth (or disparity), with the stereo-baseline changing in an increasing order. Adversarial training that goes from easy examples to harder ones produces richer losses and better models. The impact of StrAT is shown to exceed traditional data augmentation using random new views. The combination of an adversarial framework, multiview learning, and structured adversarial training produces state-of-the-art performance on unsupervised depth estimation for monocular images. The StrAT framework can benefit several problems that use adversarial training.

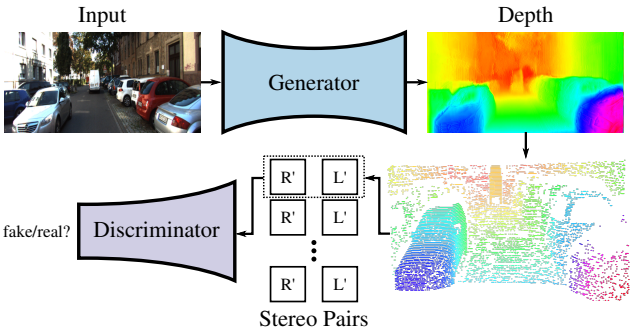

Figure 1. System overview. The generator produces stereo pairs with different baselines for every iteration using intermediate depth and the discriminator learns to disambiguate these generated images from real images. We use *structured adversarial training* in which the baseline for the stereo pairs is gradually increased over epochs.

1. Introduction

Human perception of the world in 3D has been studied in psychology, physics, mathematics and computer science. Our capacity to navigate about our surroundings relies heavily on our ability to locate ourselves with respect to the objects around us. We do this by aggregating several monocular and stereo depth cues like occlusion, motion parallax, binocular disparity, relative size and several others as discussed by Cutting and Vishton [7]. Apart from these cues, human depth perception is also based on semantic understanding of the scene. Deducing 3D shape from 2D images in geometric computer vision has been attempted using shape from X, stereo matching and structure-from-motion. These techniques infer depth from low-level depth cues but fail to incorporate scene semantics. Recently, Eigen *et al*. [8] showed that depth based on semantics can be learned by training CNNs on large datasets like KITTI [10] and NYU Depth [38]. Subsequently, several methods [6, 20, 30] have showcased the efficacy of CNNs in dense depth estimation from monocular RGB images.

These methods, however, are trained on RGB-Depth pairs. Capturing datasets with depth-aligned RGB images is expensive and requires extensive post-processing. Garg *et al*. [9] and Godard *et al*. [11] circumvent these problems by estimating dense depth (or disparity) as an intermediary task for stereo view-synthesis. They use photometric reconstruction errors to train a fully-convolutional network which estimates the adjacent stereo view corresponding to an input RGB image. Similarly, Zhou *et al*. [54] use video synthesis with depth estimation as the auxiliary task. However, these methods make two implicit assumptions: 1) brightness and colour constancy across views, and 2) Lambertian scene composition. We address these underlying assumptions in previous unsupervised depth estimation methods by proposing Structured Adversarial Training (StrAT) for generating stereo pairs from single RGB images. The adversarial learning framework (Figure 1) guides a generator network to produce realistic stereo pairs, while the discriminator learns to disambiguate between the generated and real images. Since the discriminator tries to learn the real distribution of images, the reconstruction errors incorporate the

2475-7888/18/$31.00 ©2018 IEEE
DOI 10.1109/3DV.2018.00044

Authorized licensed use limited to: INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY. Downloaded on March 07,2026 at 08:58:13 UTC from IEEE Xplore. Restrictions apply.

changes in structural content between real and synthesized images, in addition to intensity changes. Disambiguation of synthesized images from real images is a complex task. In this work, we show that using a depth-map, multiple samples at incrementally varying baselines can be generated to facilitate training of the GAN.

The success of adversarial learning in generative tasks is unchallenged. A diverse set of problems in computer vision like 3D reconstruction [13], novel view synthesis [35, 51], super-resolution [27] and image-deblurring [23] have been better solved by GANs [12]. GANs eliminate the need to find efficient loss functions for generative networks as the discriminator adapts to each task with the aim of discerning between generated and real samples. We thereby propose a generative adversarial network for the task of monocular depth estimation. To the best of our knowledge, our method is the first work which employs GANs for unsupervised depth estimation. Our main contributions in this paper are:

1. We propose a novel generative adversarial network that performs dense depth predictions for monocular RGB images using stereo-view synthesis.
2. We introduce Structured Adversarial Training (StrAT) of the GAN that goes from generating easy examples to harder ones incrementally and demonstrate its advantage.

Qualitative and quantitative evaluation shown in Section 4 demonstrate the efficacy of our method.

2. Related Work

Considering the extensive volume of literature related to depth estimation, we restrict the discussion to previous works in the area of monocular depth estimation.

Supervised Depth Estimation Most of the techniques in this area make use of depth supervision. Saxena *et al*. [36] use an MRF based approach in which the input image is over-segmented into approximately planar patches. The problem of depth estimation is then modeled as learning the 3D location and orientation of these planes. Karsch *et al*. [19] show promising results using an approach in which for a given test sample, relevant depth maps are extracted from a database of RGBD pairs which are warped according to the given sample. The warped depth maps are then refined using optical flow and motion estimation. This method, however, requires the entire database during inference. Ladicky *et al*. [25] argue that semantic segmentation can be used for inferring the relative size (a strong cue for depth estimation) of the objects in the scene. In addition to these methods, CNN based techniques have also shown promise in the past few years. Eigen *et al*. [8] show the first prominent work in the area. Their approach learns pixel-aligned depth from monocular images and unlike previous methods [37], it does not use hand-crafted features. Liu *et al*. [29] propose a CNN-based learning scheme which learns unary and pairwise potentials for a CRF model. With the introduction of fully convolutional networks by Long *et al*. [31], Mayer *et al*. [33] propose an FCN for single image depth estimation and optical flow, which has subsequently resulted in several improvements in the area [11, 28, 32, 46, 50, 54].

Unsupervised Depth Estimation Capturing ground-truth depth maps is expensive and time consuming. An alternate strategy to train CNNs for depth estimation is to use multi-view supervision as a proxy for depth supervision. Garg *et al*. [9] use stereo view supervision, in which the left view of a stereo pair is warped into the right view by learning dense disparities as an intermediary task. They make Taylor approximations to engineer differentiable loss functions. Godard *et al*. [11] instead use bilinear sampling and add a consistency check between the disparities learned for the left view and the right view. Similarly, Xie *et al*. [48] introduce stereo view synthesis as a problem of learning probability distributions over disparities. Augmenting view supervision with depth supervision further improves the accuracy of depth estimation as shown by Kuznietsov *et al*. [24].

Video synthesis has also been used as an alternative to stereo synthesis. For stereo synthesis, the baseline is assumed to be fixed throughout the training set, most frameworks implicitly use camera-pose supervision. For video synthesis, the additional challenge is that the transformation between two consecutive frames is variable. Zhou *et al*. [54] use a pose network in addition to a depth network and use video reconstruction errors to train them together. Simultaneously, Vijayanarasimhan *et al*. [45] learn depth and pose change with various degrees of depth supervision. In spite of using only view reconstruction errors, these approaches showcase better results than a depth supervised method [8]. In an attempt to add more geometric constraints, Yang *et al*. [50] add a consistency loss function for surface normals inferred from the depth and input RGB image. Alternatively, Mahjourian *et al*. [32] use ICP-based [5] loss on the 3D point clouds generated using a depth network. Furthermore, Zhan *et al*. [52] integrate both stereo and video synthesis in their method (like [50]) along with a deep-feature loss.

There also have been attempts on using other alternatives for supervision such as camera aperture [40] and light fields [41] for dense stereo. But stereo supervision [11] still remains as the most accurate and memory-efficient technique. With the introduction of stereo cameras in several modern smartphones, capturing stereo images is almost as easy as capturing monocular images. Consequently, we make use of only stereo-supervision for our method.

Authorized licensed use limited to: INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY. Downloaded on March 07,2026 at 08:58:13 UTC from IEEE Xplore. Restrictions apply.

Most of the view supervision based methods use loss functions for low-level features like pixel intensities. These methods inadvertently make assumptions about the scene being composed of only Lambertian and densely-textured surfaces. The computed errors are highly local and fail at capturing high-level semantics *e.g.* objects like poles and trees must remain rigid across views. These errors *viz.* Structural Similarity (SSIM) and $\ell 1$, do not account for the entire context of a scene and hence can result in unreliable depth estimates (Figure 2). We propose a generative adversarial network in which the discriminator considers a larger context while computing the loss between a generated and a target image. Concurrent to our work, Atapour-Abarghouei and Breckon [3] also use GANs for monocular depth estimation, but with a supervised approach and a supplementary dataset. Our method is unsupervised as it requires only stereo images for training and it even performs better than most depth-supervised techniques.

3. Depth using Adversarial Training

Point-to-point correspondence between the left view and the right view of a stereo pair can be used to estimate dense depth maps as:

$$D^i = \frac{Bf}{d^i}, \tag{1}$$

where D^i is the depth for the point i with disparity d^i, B is the baseline for the stereo camera and f is the focal length. With a known dense disparity map, a stereo pair can be generated from the image corresponding to the disparity map. Inspired by this, the naïve way to compute depth would be to generate the adjacent stereo-view from a single image and use stereo-matching algorithms to estimate the depth of the scene, in a two-step approach. We use a one-step approach instead, in which a dense disparity map is generated from a single image, which is in-turn used to create the adjacent stereo view. The errors in the generated disparity map are computed using reconstruction errors between the generated and the ground-truth stereo view. This is similar to novel view synthesis using appearance flow [55]. In case of rectified stereo pairs, the appearance flow becomes one dimensional due to epipolar constraints. The magnitude of the flow for every pixel represents its corresponding disparity and dense depth can be inferred according to Equation 1.

Our goal is to generate a stereo pair (L'_i, R'_i) from a single image L_i, assuming it to be the left view without loss of generality. To this end, we train a generator network as a feed-forward CNN G_{θ_G}, where θ_G are the parameters of the network. The generator takes L_i as the input and generates a pair of disparities (D_i^L, D_i^R). Subsequently, a bilinear sampler (S) [16] generates the stereo pair as $L'_i = S(D_i^L, R_i)$ and $R'_i = S(D_i^R, L_i)$. Note that (L_i, R_i) pairs are present only during training. During inference, a forward pass

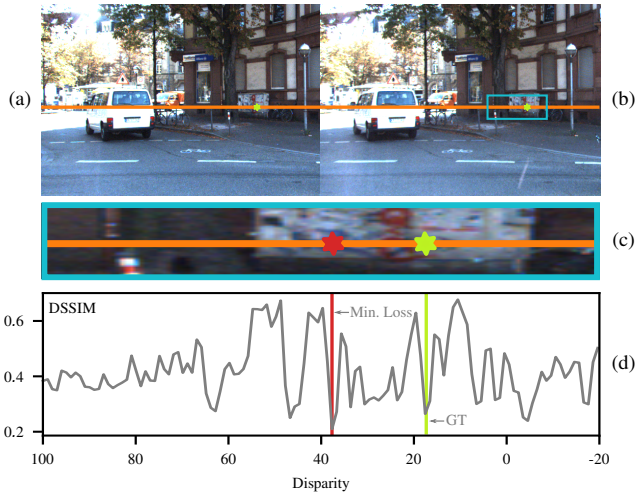

Figure 2. Unreliable stereo-matching using structural similarity (SSIM). (a) and (b) are a stereo pair from KITTI [10]. The line in orange denotes the epipolar line. The green stars show ground-truth correspondence. The red star shows the estimated match using SSIM. The search region for stereo matching is highlighted in blue. DSSIM $= \frac{1-\text{SSIM}}{2}$

through G with L_i as the input is sufficient to produce dense depth maps from D_i^L using Equation 1. For training pairs $\{(L_i, R_i) \mid i = 1 \dots N\}$, we solve:

$$\hat{\theta}_G = \arg\min_{\theta_G} \frac{1}{N} \sum_i E^R(R'_i, R_i) + E^L(L'_i, L_i), \tag{2}$$

where E^L and E^R are linear combinations of a set of loss functions defined in later parts of this section. For the remainder of this paper, we restrict the discourse to generation of only right views R'_i.

Previous methods for depth estimation based on view-synthesis [11, 54] rely on primitive reconstruction loss functions like mean-squared error and SSIM [47]. These pixel-wise error measurements fail to capture the perceptual quality of a generated image as shown by Zhang *et al.* [53]. As shown in Figure 2, a pixel-based loss like structural dissimilarity, when used for stereo matching, produces unfavourable loss surfaces with multiple local minima and a global minimum far from the ground truth disparity. Pixel-based losses are also unsuitable for scenes with non-Lambertian surfaces, which is not the case with perceptual loss functions.

One way to capture the perceptual loss between two images, is to use feature maps from a pre-trained network like VGG19 [39], as proposed by Johnson *et al.* [17]. However, a network like VGG19 trained on a classification task does not capture high-frequency details required for stereo matching. We instead train a discriminative network which adapts to our particular task by learning to distinguish a generated image from a real image.

Authorized licensed use limited to: INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY. Downloaded on March 07,2026 at 08:58:13 UTC from IEEE Xplore. Restrictions apply.

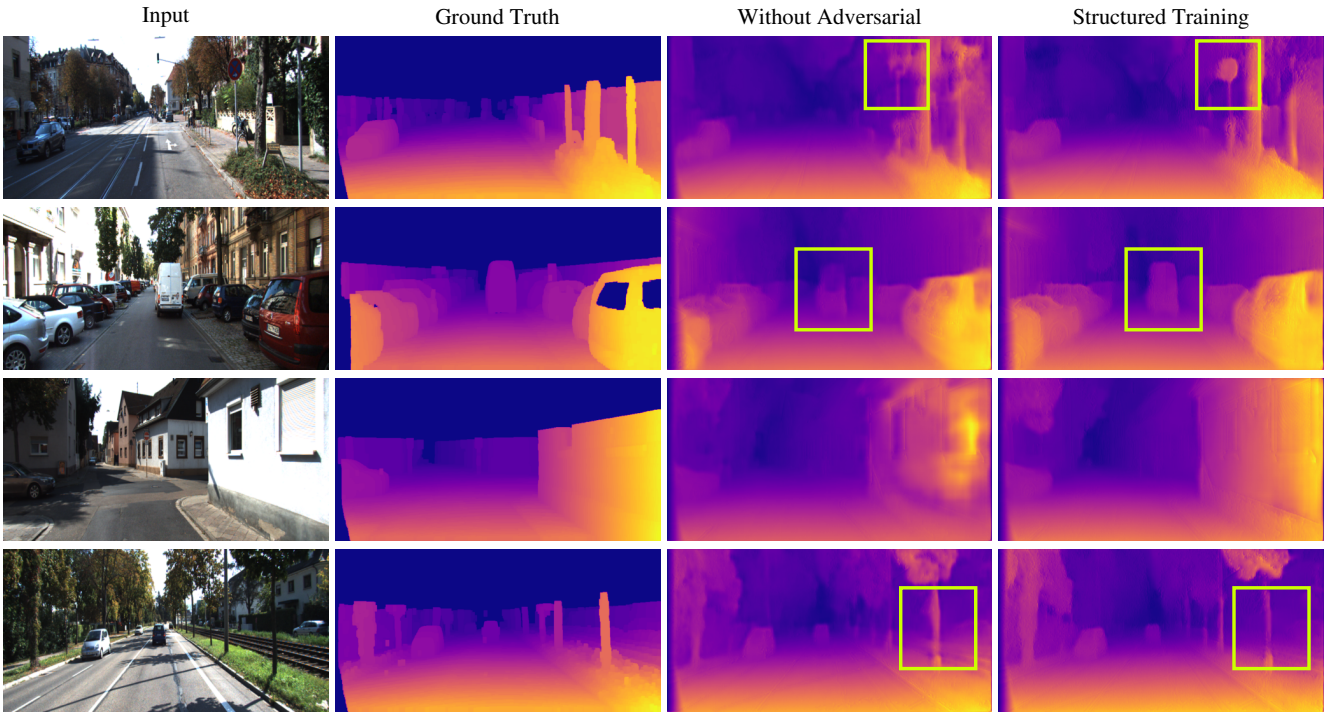

Figure 3. Qualitative comparison between our model with and without adversarial learning. The model using adversarial loss performs better for objects with low-texture density and for thin structures like poles and trees. The major differences are inside the boxed regions. For better visualization, we dilate the ground-truth depth maps which are Velodyne laser projections and hence sparse.

3.1. Adversarial Learning

We propose an adversarial framework for view synthesis, which eliminates the need to find an appropriate loss function that captures the perceptual quality of the generated view. We use a discriminator network which learns to predict whether the synthesized view belongs to the manifold of real views without relying on pixel-by-pixel matching. In the proposed framework, the generator G_{θ_G} learns to synthesize the adjacent right view for an input image using the generated disparity D_i^R. The discriminator C_{θ_C} tries to distinguish the generated right view from the real right views in the training set, as it competes with the generator. This is done by optimizing the following objective in an alternating manner:

$$\mathcal{L}_{adv}^{R} = \min_{\theta_G} \max_{\theta_C} \mathbb{E}_{L\sim p_G(L)}[\log(1 - C_{\theta_C}(R_i')] + \mathbb{E}_{R\sim p_{train}(R)}[\log\ C_{\theta_C}(R)] \tag{3}$$

Equation 3 represents the vanilla GAN [12] objective for generating right view R_i' from left view L_i. It is to be noted that for view synthesis of outdoor scenes like in KITTI [10] the transformation of objects across views is mostly rigid. Our goal is to learn a generator which is aware of such structural content in the scene and other semantics, *e.g.*, thin structures like poles remain rigid across all views.

In a naïve training procedure, the generator produces right views with a fixed baseline B. In the early stages of training, since the estimated disparities are quite erroneous, using them at full scale (with baseline B) will generate right views which are easily distinguishable from the original training set. As a result, the discriminator dominates the training procedure and hence equilibrium is not achieved for the adversarial objective (Equation 3). We resolve this by adding more consistency constraints by using multiple disparate views corresponding to the same disparity map.

3.2. Multi-View Synthesis

The generator network produces a dense depth-map which is a 2.5D representation of the scene. As a result, given a camera-transformation matrix, a novel view can be generated corresponding to the transformation. With more views, we can add additional consistency constraints to the depth.

For accurate depth estimation, the generated views should be *real-looking* and consistent with each other. Inspired by this idea, we can train our framework to generate multiple views of the scene corresponding to an input image, using the dense-depth map from the generator. The views can be generated using the dense-depth map obtained from the generator. The task of the discriminator remains to discern these views from real views present in the training

Authorized licensed use limited to: INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY. Downloaded on March 07,2026 at 08:58:13 UTC from IEEE Xplore. Restrictions apply.

data. This results in a min-max zero-sum game in which the generator and the discriminator are the two agents.

In case of a complex scene, for generating views from 2.5D with severe changes in camera pose, an additional step of hole-filling becomes necessary. The task of the discriminator in such cases becomes fairly easy and equilibrium for the min-max game is not achieved, as we find in our experiments. To generate views with small pose changes, we restrict the pose transformation as a translation along the direction of the epipolar lines. This is equivalent to generating right views with fractional baselines:

$$R'_{i,b} = S\big(D_i^R \cdot \frac{b}{B}, L_i\big), \tag{4}$$

where $R'_{i,b}$ corresponds to baseline b. For the sake of simplicity, we take R'_i as the right view corresponding to baseline B. Note that the training data consists of stereo pairs with baseline B only. For every training iteration, a right view is generated with a chosen baseline b. We show that the choice of b is crucial for training and propose a structured training scheme which varies b in a sequential and organized manner.

3.3. Structured Adversarial Training

Inspired by the work on curriculum learning by Bengio *et al*. [4], we argue that by training our framework with right views in a meaningful order helps the discriminator learn in a more structured manner. We gradually increase the baseline such that the impact of the erroneous disparities in the initial learning stages is mitigated. In other words, b is linearly varied from 0 to B over epochs. The incremental nature of this scheme allows the training to first discover low-level mismatches like incorrect textures, and then shift attention to high-level content changes. In the early stages of training, the task is substantially simpler as there is minimal loss of information between stereo views with small baseline. As we increase the baseline b incrementally, the task becomes harder as the effect of erroneous disparities becomes pronounced. This procedure is conceptually similar to the recent work by Karras *et al*. [18], in which they progressively train GANs to first generate low-resolution images and then high-resolution images.

We find that linearly varying b performs better than just randomly sampling b from $[0, B]$, which itself is better than using a fixed baseline B. Training with random baseline can be considered as data augmentation to regularize the discriminator. However, our experiments indicate that not only does the learning benefit from stereo-views with multiple baselines, but also from structured training. We find in our experiments that linearly varying b results in much more stable training compared to other training schemes.

3.4. Training the Generator

In this section, we describe the loss functions used to train the generator. We use a combination of adversarial loss and photometric loss to achieve the desired outcome.

Adversarial Loss The adversarial objective influenced by the parameters θ_G of the generator is minimized to train the generator:

$$\mathcal{L}^R_{G.adv} = -\log C_{\theta_c}(R'_{i,b}) \tag{5}$$

We optimize $-\log C_{\theta_c}(R'_{i,b})$ instead of $\log[1 - C_{\theta_c}(R'_{i,b})]$ for better gradient behaviour. The adversarial loss encourages the generator to generate right views which reside on the manifold of ground-truth right views and as a result, fool the discriminator. The discriminator is not conditioned by L_i and hence only learns to perceptually disambiguate R'_i from the distribution of real images and not whether R'_i corresponds to L_i. As a result, using the adversarial loss independently will lead to $R'_i = L_i$ as L_i belongs to the distribution of real images. To avoid this problem of mode collapse, we couple the adversarial objective with other metrics like $\ell 1$ and SSIM loss used in several prior works [11, 45, 50, 54].

Photometric Reconstruction Loss We use photometric losses to incentivize the generator to create images which are pixel-wise similar to right views corresponding to input left views. The SSIM loss for right views $\{R'_i \mid i = 1 \dots N\}$ is:

$$\mathcal{L}^R_{\text{SSIM}} = \frac{1}{w \times h} \sum_x \sum_y \frac{1 - \text{SSIM}\{R'_i(x,y), R_i(x,y)\}}{2}, \tag{6}$$

where w is the width and h is the height of R_i. $R_i(x,y)$ is a patch at (x,y) with an empirically determined patch size of 3×3. Additionally, for local matching we use $\ell 1$ loss:

$$\mathcal{L}^R_{\ell 1} = \|(R'_i - R_i)\|_1 \tag{7}$$

Left-Right Consistency To ensure coherence between left-disparity D_i^L and right-disparity D_i^R, a consistency check is introduced [11]. The left-disparity is taken as the left view of a stereo pair and is warped into its corresponding right view (here, right-disparity $D_i^{R'}$) as $D_i^{R'} = S(D_i^R, D_i^L)$. An $\ell 1$ penalty is used for the projected and generated right-disparities:

$$\mathcal{L}^R_{CR}(D_i^R, D_i^L) = \left\| \left(D_i^{R'} - D_i^R \right) \right\|_1 \tag{8}$$

Relating equations 2,5,6,7 and 8, the final objective for the generator to generate right views from left views is:

Authorized licensed use limited to: INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY. Downloaded on March 07,2026 at 08:58:13 UTC from IEEE Xplore. Restrictions apply.

	Error (lower the better)				Precision (higher the better)		
	Abs. Rel	Sq. Rel	RMSE	log RMSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Yang *et al*. [50] *	0.172	1.559	6.794	0.252	0.744	0.910	0.969
Godard *et al*. [11]	0.124	1.388	6.125	0.217	0.841	0.936	0.975
Ours – No $\mathcal{L}_{adv}$	0.118	1.124	5.916	0.211	0.835	0.936	0.975
Ours – Fixed Baseline	0.115	1.154	5.766	0.205	0.846	0.944	0.979
Ours – Random Baseline	0.112	1.089	5.636	0.201	0.850	0.946	0.980
Ours – StrAT	**0.108**	**1.019**	**5.551**	**0.195**	**0.856**	**0.950**	**0.981**

Table 1. Results on KITTI [10] 2015 split. Comparison of our model with previous unsupervised methods. Our model with structured adversarial training (StrAT) scheme performs the best across all metrics. (* – trained on monocular videos)

$$E^R = \lambda_1 \mathcal{L}^R_{G.adv} + \lambda_2 \mathcal{L}^R_{\text{SSIM}} + \lambda_3 \mathcal{L}^R_{\ell 1} + \lambda_4 \mathcal{L}^R_{CR}(D^R_i, D^L_i), \tag{9}$$

where λ_{1-4} are empirically chosen, the details for which are mentioned in Section 4.1. All the discussed loss functions are mirrored to get a similar objective for E^L.

3.5. Network Architecture

The generator is a fully convolutional network with an encoder and a decoder. It is similar to the DispNet architecture proposed by Mayer *et al*. [33]. We use a ResNet-50 [14] based encoder for its large receptive field inspired by [26], and it is composed of four residual blocks. The encoder is followed by a convolutional layer with 1×1 kernel and then by a decoder. The decoder is composed of multiple transposed convolutions to upsample the encoded feature map. We use skip-connections to ensure that there is minimal information loss due to downsampling and that the gradients flow smoothly. Unlike the previous non-linearity layers which use ELU [43], our final layer is followed by ReLU [34] to ensure that the resultant disparities are positive. The output of the generator is a $512 \times 256 \times 2$ matrix with concatenated left and right disparity maps.

The discriminator is similar to PatchGAN [15] with a series of convolutional layers. Except the last layer, all of them are followed by Instance Normalization [44], Leaky ReLU [49] with $\alpha = 0.2$ and Dropout [42] regularization with keep-probability as 0.8. The last layer is followed by sigmoid activation. The input to the discriminator is a random patch of 256×256 from the generated and ground-truth images. The PatchGAN generates a 7×7 grid in which every cell indicates whether the input image patch in its receptive field is real or fake.

4. Experiments

We show thorough quantitative and qualitative evaluation by analyzing the performance of our method for monocular depth estimation for the training and test splits discussed in Section 4.2. We compare our work with prior methods on a standard benchmark and report promising results. To enable reproducibility, we also list various implementation details. Ablation studies on the adversarial loss illustrate its validity. We also highlight the generalization of our model on unseen data. Finally, we show the performance of depth estimation individually on a few object categories using semantic segmentation.

4.1. Implementation Details

Our models are implemented using the Tensorflow [1] deep learning framework. Training is performed on two NVIDIA 1080 Ti GPUs. It is executed for 50 epochs with $b = (0.1 + \frac{e}{50} \times 0.9)B$, where b is the baseline for epoch e and B is the baseline for the stereo pairs in the training set. The initial learning rate is 0.0001, which is linearly decreased to zero in the second half of the training. For gradient descent optimization, we use Adam [22] optimizer with the settings as $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. For loss balancing, we using $\lambda_1 = 0.01$, $\lambda_2 = 1.0$, $\lambda_3 = 0.1$ and $\lambda_4 = 1.0$ in Equation 9. Following [11] and [54], the photometric losses are calculated at four different scales and added together with equal weights. To make the discriminator more robust against colour and intensity variation, we augment discriminator inputs by randomly scaling the colour channels individually with factors from 0.8 to 1.2.

4.2. Datasets

We evaluate our model primarily on the KITTI [10] dataset, having 42,382 rectified stereo pairs from 61 videos captured from a camera rig mounted on the roof of a car. All the images have a resolution of 1242×375, which we downsample to 512×256 for memory efficiency. We perform the evaluation at full-resolution by resizing the depth maps using bilinear interpolation. To compare our work with prior work on depth estimation, we use two commonly used train-test splits.

KITTI Split Out of the 42,382 stereo pairs, the training set comprises of 30,159 images from 33 videos. The test set contains 200 images with sparse disparities from 28 videos.

Authorized licensed use limited to: INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY. Downloaded on March 07,2026 at 08:58:13 UTC from IEEE Xplore. Restrictions apply.

Semantic Categories	Methods		
	[11]	Ours - Photo.	Ours - Adv
Flat	2.165	2.376	2.149
Human	3.190	2.763	2.770
Vehicle	3.690	3.968	3.507
Construction	9.713	9.936	9.364
Object	12.763	11.789	11.537

Table 2. RMS errors for semantically categorized objects. From top to bottom, the categories are arranged in increasing order of textural uncertainty. Our model with adversarial loss is less error prone for objects like sign-boards, traffic lights, poles, *etc*., than our model trained with only photometric losses.

We split the training set further in validation and train splits. The validation set contains 10 images randomly sampled from every video in the training set. The remaining 29,829 images make the final training set. For evaluation we use the metrics proposed by Eigen *et al*. [8] as shown in Table 1. Our method with structured adversarial learning estimates more accurate disparities compared to previous unsupervised methods.

Eigen Split Table 3 shows comparisons with other methods on the train-test split proposed in [8]. The test set in this split comprises of 697 images with corresponding sparse depth maps. These images are present in 29 scenes and the training set spans the rest of the 32 scenes. We use the same train-val split strategy that we use for KITTI split, to form a training set with 23,168 images and a validation split with 320 images.

4.3. Comparisons with other Methods

We show quantitative comparison with previous monocular depth estimation methods in Table 1 for the KITTI split and Table 3 for the Eigen split. For the Eigen split, the evaluation is performed inside the crop regions proposed in [9]. We use the same evaluation metrics used in previous works [3, 9, 11, 24, 45, 50, 54]. Our method performs better than other unsupervised methods [11, 45, 50, 54] and is comparable with supervised methods [3, 24] both in terms of precision and errors. We use the metrics defined in [8] and calculate the errors in the depth space. Since our network predicts disparities and not depth, it could lead to precision issues.

4.4. Ablation Study

We perform an ablation study to establish the benefit of using adversarial learning for the problem of monocular depth estimation. The results of the study are shown in Table 1. We evaluate the performance of our model with $\lambda_1 = 0$, *i.e*. without adversarial loss (shown as *Ours - No* $\mathcal{L}_{adv}$). We compare its performance with the model trained with adversarial loss using three baseline choices – fixed, random and structurally varying. Our model with structured adversarial training performs the best across all methods over all metrics. The model trained with fixed baseline training scheme is comparable to the model trained without adversarial loss. This indicates that training without using stereo-pairs with different baselines is ineffective which is in accordance with our hypothesis in Section 3.3.

We show qualitative comparison between our model without adversarial learning and with structured adversarial training in Figure 3. The structured training scheme generates better depth for thin structures like poles and low texture density surfaces like a white wall. This can become noteworthy for challenging datasets.

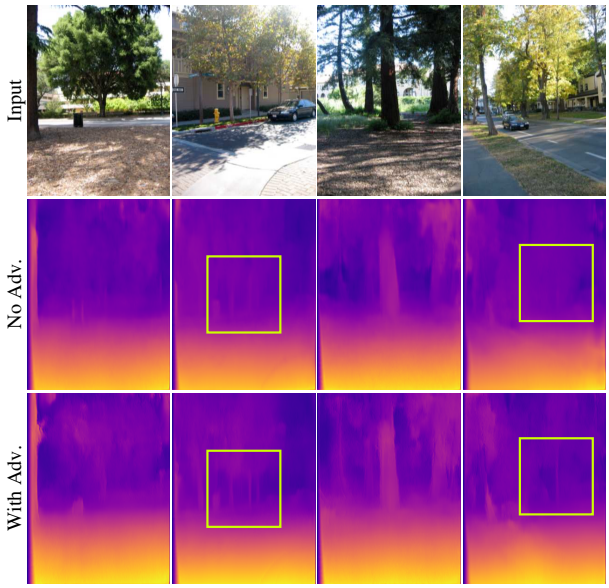

Figure 4. Generalization results. The model is trained only on KITTI [10] and the test images are from Make3D [37]. The model trained with adversarial loss performs better than the model trained on photometric loss terms. The differences are highlighted in the boxed regions. Adversarial training has helped in segmenting the objects better in the scene.

4.5. Semantic Errors

To evaluate the depth estimation performance on specific object categories, we make use of semantic segmentation maps released for the KITTI test set [2]. We classify the pixels in every image into five object categories as shown in Table 2. The RMSE for flat surfaces like road and pavements is the lowest and for objects like sign-boards, traffic-lights, *etc*., it is the highest. The structured training model achieves positive results for all the semantic categories. Note that the errors for the *constructions* category is higher than the errors for the *human* category. We believe that the errors are proportional to the uncertainty in the appearance of the objects in terms of their texture, shape and colour.

Authorized licensed use limited to: INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY. Downloaded on March 07,2026 at 08:58:13 UTC from IEEE Xplore. Restrictions apply.

	Error (lower the better)				Precision (higher the better)		
	Abs. Rel	Sq. Rel	RMSE	log RMSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen *et al.* [8] – Fine *	0.203	1.548	6.307	0.282	0.702	0.890	0.958
Liu *et al.* [30] *	0.202	1.614	6.523	0.275	0.678	0.895	0.965
Kuznietsov *et al.* [24] *	0.122	0.763	4.815	0.194	0.845	0.957	0.987
Amir *et al.* [3] †	0.110	0.929	4.726	0.194	0.923	0.967	0.984
Zhou *et al.* [54]	0.183	1.595	6.709	0.270	0.734	0.902	0.959
Yang *et al.* [50]	0.182	1.481	6.501	0.267	0.725	0.906	0.963
Godard *et al.* [11]	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Zhan *et al.* [52]	0.135	1.132	5.585	0.229	0.820	0.933	**0.971**
Ours – StrAT	**0.128**	**1.019**	**5.403**	**0.227**	**0.827**	**0.935**	**0.971**
Zhou *et al.* [54]	0.201	1.391	5.181	0.264	0.696	0.900	0.966
Garg *et al.* [9]	0.169	1.080	5.104	0.273	0.740	0.904	0.962
Godard *et al.* [11]	0.140	0.976	4.471	0.232	0.818	0.931	0.969
Zhan *et al.* [52]	0.128	0.815	4.366	0.225	0.818	0.937	0.973
Ours – StrAT	**0.122**	**0.768**	**4.095**	**0.214**	**0.842**	**0.943**	**0.974**

Table 3. Results on KITTI [10] train-test split proposed by Eigen *et al.* [8]. All the methods except [8] are evaluated on the crop region used in [9]. The results are categorized according to the maximum depth cap (80m – upper half, 50m – lower half). * – trained with additional depth supervision. [3] † is trained on a synthetic dataset with depth supervision. Our method with structured adversarial training (StrAT) performs better than all other unsupervised techniques and is comparable with the supervised ones.

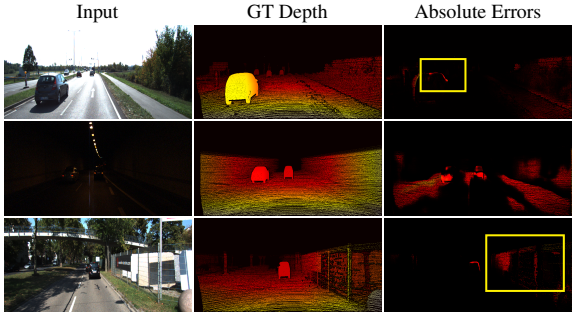

Figure 5. Failure cases. Our model estimates incorrect depth for object boundaries, low-intensity regions and objects with high textural uncertainty. Most of the issues can be resolved with more training data.

4.6. Generalization

We show generalization of our approach in Figure 4, in which our model, trained with and without adversarial learning are compared. Both the models are trained on KITTI [10] and the test samples are from Make3D [37]. Even though the appearance of the objects in Make3D is different from the ones in KITTI, adversarial training has helped in capturing more variance in object appearance.

4.7. Limitations

Just like other stereo-supervision methods [9, 11], our model produces erroneous depth for object boundaries. This is because stereo-matching implicitly assumes that the scene is devoid of occlusions. Also, generating adjacent stereo-views from a single image using only disparities is over-constrained. We show this problem with stereo-supervision methods and more failure cases in Figure 5. Unsurprisingly, our method also fails for low-intensity regions. This is solely because of information loss in these regions and is not a drawback of our method. As shown quantitatively in Table 2 and qualitatively in Figure 5, our model generates erroneous depth for the *constructions* category. This limitation can be attributed to epistemic uncertainty [21] and it can be resolved with more training data.

5. Conclusion and Future Work

We propose a generative adversarial network to improve unsupervised depth estimation by capturing an overall perceptual loss instead of pixel-by-pixel error measures. We also introduce multi-view training and a structured adversarial training framework to ensure the network learns systematically from easy examples to hard ones for this problem. We show quantitatively that adding structure into the process of adversarial training can substantially improve depth estimation.

The intermediate view synthesis step lends itself to the generation of such examples by varying the underlying baseline parameter. Several dense estimation tasks like scene-refocusing, optical flow and depth estimation also have similar parametric structure to generate additional training samples to be used in a structured training framework, going beyond what is available in the training set. Our structured training framework will easily generalize to such problems. We propose to explore these in the future.

Authorized licensed use limited to: INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY. Downloaded on March 07,2026 at 08:58:13 UTC from IEEE Xplore. Restrictions apply.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, 2016.

[2] H. A. Alhaija, S. K. Mustikovela, L. Mescheder, A. Geiger, and C. Rother. Augmented reality meets deep learning for car instance segmentation in urban scenes. In *BMVC*, 2017.

[3] A. Atapour-Abarghouei and T. Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation. In *CVPR*, 2018.

[4] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML*, 2009.

[5] P. J. Besl and N. D. McKay. A method for registration of 3-d shapes. In *PAMI*, 1992.

[6] Y. Cao, Z. Wu, and C. Shen. Estimating depth from monocular images as classification using deep fully convolutional residual networks. In *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[7] J. E. Cutting and P. M. Vishton. Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In *Perception of space and motion*, pages 69–117. Elsevier, 1995.

[8] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*, 2014.

[9] R. Garg, V. K. BG, G. Carneiro, and I. Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *ECCV*, 2016.

[10] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012.

[11] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014.

[13] J. Gwak, C. B. Choy, M. Chandraker, A. Garg, and S. Savarese. Weakly supervised 3d reconstruction with adversarial constraint. In *3DV*, 2017.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[16] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *NIPS*, 2015.

[17] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[18] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018.

[19] K. Karsch, C. Liu, and S. B. Kang. Depth extraction from video using non-parametric sampling. In *ECCV*, 2012.

[20] K. Karsch, C. Liu, and S. B. Kang. Depth transfer: Depth extraction from videos using nonparametric sampling. In *PAMI*, 2016.

[21] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NIPS*, 2017.

[22] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[23] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *CVPR*, 2018.

[24] Y. Kuznietsov, J. Stückler, and B. Leibe. Semi-supervised deep learning for monocular depth map prediction. In *CVPR*, 2017.

[25] L. Ladicky, J. Shi, and M. Pollefeys. Pulling things out of perspective. In *CVPR*, 2014.

[26] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *3DV*, 2016.

[27] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2016.

[28] R. Li, S. Wang, Z. Long, and D. Gu. Undeepvo: Monocular visual odometry through unsupervised deep learning. In *ICRA*, 2018.

[29] F. Liu, C. Shen, and G. Lin. Deep convolutional neural fields for depth estimation from a single image. In *ICCV*, 2015.

[30] F. Liu, C. Shen, G. Lin, and I. Reid. Learning depth from single monocular images using deep convolutional neural fields. In *PAMI*, 2016.

[31] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[32] R. Mahjourian, M. Wicke, and A. Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *CVPR*, 2018.

[33] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016.

[34] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.

[35] K. Regmi and A. Borji. Cross-view image synthesis using conditional gans. In *CVPR*, 2018.

[36] A. Saxena, S. H. Chung, and A. Y. Ng. Learning depth from single monocular images. In *NIPS*, 2006.

[37] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. In *PAMI*, 2009.

[38] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[39] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[40] P. P. Srinivasan, R. Garg, N. Wadhwa, R. Ng, and J. T. Barron. Aperture supervision for monocular depth estimation. In *CVPR*, 2018.

[41] P. P. Srinivasan, T. Wang, A. Sreelal, R. Ramamoorthi, and R. Ng. Learning to synthesize a 4d rgbd light field from a single image. In *ICCV*, 2017.

Authorized licensed use limited to: INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY. Downloaded on March 07,2026 at 08:58:13 UTC from IEEE Xplore. Restrictions apply.

[42] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. In *JMLR*, 2014.

[43] L. Trottier, P. Gigu, B. Chaib-draa, et al. Parametric exponential linear unit for deep convolutional neural networks. In *IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2017.

[44] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. In *CoRR*, volume abs/1607.08022, 2016.

[45] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki. Sfm-net: Learning of structure and motion from video. In *arXiv preprint arXiv:1704.07804*, 2017.

[46] C. Wang, J. M. Buenaposada, R. Zhu, and S. Lucey. Learning depth from monocular videos using direct methods. In *CVPR*, 2018.

[47] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. In *IEEE Transactions on Image Processing*, 2004.

[48] J. Xie, R. Girshick, and A. Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *ECCV*, 2016.

[49] B. Xu, N. Wang, T. Chen, and M. Li. Empirical evaluation of rectified activations in convolutional network. In *ICML Deep Learning Workshop*, 2015.

[50] Z. Yang, P. Wang, W. Xu, L. Zhao, and R. Nevatia. Unsupervised learning of geometry with edge-aware depth-normal consistency. In *AAAI*, 2018.

[51] X. Yin, H. Wei, X. Wang, Q. Chen, et al. Novel view synthesis for large-scale scene using adversarial loss. In *arXiv preprint arXiv:1802.07064*, 2018.

[52] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *CVPR*, 2018.

[53] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[54] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017.

[55] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros. View synthesis by appearance flow. In *ECCV*, 2016.

Authorized licensed use limited to: INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY. Downloaded on March 07,2026 at 08:58:13 UTC from IEEE Xplore. Restrictions apply.